単語を共有する文書グラフを用いた文書分類

井田 龍希　三輪 誠　佐々木 裕
豊田工業大学
{sd18006,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

文書分類において，文書のテキスト情報を BERT でエンコードして分類する手法が主流である．一方で，文書とそれが共有する単語を節点とした文書グラフを利用し文書間の関係を考慮した文書分類が提案されている．それぞれの手法は，文書の異なる観点を捉えていると考えられ，両方を活かした文書分類ができれば精度の向上が期待できる．そこで，本研究では BERT を利用して文書グラフの文書・単語節点表現を初期化し，両者の利点を効果的に利用する手法を提案する．実験では，本手法による精度向上を確認し，文書内のテキスト情報と文書グラフを利用することの有用性を示した．また，文書分類における文書グラフの単語節点の影響も調査した．

1　序論

文書分類においては，文書のテキスト情報のみを利用して分類を行う手法が主流である．近年，大量のデータで事前学習をした BERT (Bidirectional Encoder Representations from Transformers) [1] を用いることで，少ないデータで微調整をするだけでタスクに特化したモデルが作成でき，大量のテキストデータによる事前学習と文書内のテキスト情報を考慮した表現によって高い精度が達成されている．

一方で，文書間の関係を表した文書グラフを利用した文書分類が提案されている．Yao ら [2] は，文書と単語を節点とし，文書節点とその文書に出現する単語節点の間・関連度の高い単語間それぞれに辺を張った文書グラフを用いて，文書分類を行っている．この文書グラフにより，共通した単語を持つ文書節点がその単語節点を介してつながり，文書間の関係を考慮した文書分類ができる．

それぞれの手法は文書内のテキスト情報と文書間の関係を表す文書グラフという文書の異なる観点を捉えていると考えられるので，この両方を活かした文書分類ができれば，より高い精度が達成できると期待できる．これまで，文書節点の表現に BERT を利用した文書グラフを用いて文書分類を行う手法 [3] が提案されているが，単語は文書をつなぐ節点としてのみ利用されており，文書グラフとテキスト情報のつながりが文書節点のみになってしまっているので，両者が十分に連携できているとは言えない．

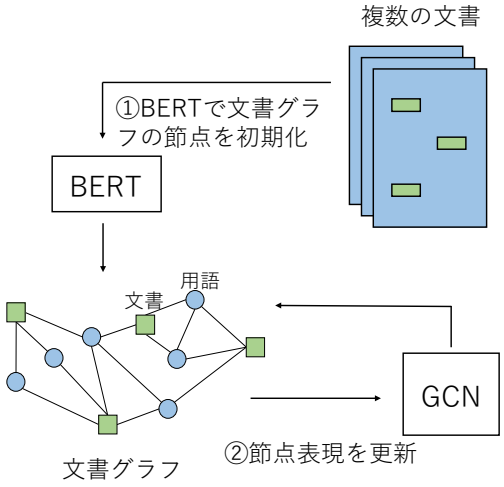

図 1　提案手法の概要

そこで，本研究では，BERT を利用して文書グラフの全節点にテキスト情報を追加し，テキスト情報と文書グラフを密接に連携させ，両者をより効果的に利用する文書分類の実現を目指す．

2　関連研究

文書分類には文書のテキスト情報のみを使用する手法と文書間の関係を考慮する手法がある．

文書のテキスト情報のみを使う手法では BERT を用いた手法が高い精度を達成している．この手法では BERT でテキストをエンコードして得られた表現を使って文書分類を行う．大量のデータで事前学習した BERT を少ないデータで文書分類に特化するように学習させることで，文書全体の文脈を考慮した表現が得られ高い精度が達成できる．

文書間の関係を考慮する手法では，文書グラフを用いた手法があり，Yao らは文書と単語を節点と

し，文書節点とその文書に出現する単語節点の間に TF-IDF 値で重み付けした辺・PMI の値によって関連度が高いとみなされた単語節点間の辺の 2 種類の辺を張った文書グラフを用いた文書分類を提案した．Yao らは節点に接続する辺を通して周りの節点からの情報をその節点に集約するグラフ畳み込みネットワーク（GCN; Graph Convolutional Network) [4] を用いて，文書グラフの節点表現をグラフ構造を考慮するように更新し，分類に利用している．さらに，BertGCN[3] では，文書節点の表現に BERT を利用し，文書グラフに文書のテキスト情報を追加する手法を提案し，高い精度を達成している．

3 提案手法

本研究では，テキスト情報と文書グラフの情報の両方を効果的に利用する文書分類モデルを提案する．両方を利用した BertGCN[3] に対して，本研究では，より密接な情報の連携を目指して，文書・単語節点の両方を BERT で初期化した文書グラフを用いて文書分類をする．ここでは，この文書グラフの作成方法について説明した後に文書分類に用いるモデルについて説明したのちに，学習と予測について説明する．

3.1 モデル

文書グラフの作成の概要を図 2 に示す．本研究では Yao らが提案した文書と単語を節点とする文書グラフとそのモデルを基盤として，文書・単語の両方の節点に BERT の表現を追加して使用する．まず，文書のテキスト情報を BERT でエンコードして各トークンの表現を得る．BERT の出力の [CLS] トークンは文全体を表す表現になっているので，[CLS] トークンの表現を文書の初期表現として，TextGCN の表現である one-hot ベクトルに結合する．この際，BERT の表現のスケールを one-hot ベクトルのスケールと合わせるために，BERT の表現は L2 正規化したものを用いる．単語の表現についても同様に BERT の表現を追加する．単語の表現については，BERT ではサブワードをトークンの単位として表現しているため，サブワードの表現を平均としたものを利用する．今回使用する文書グラフでは，同じ表層の単語が複数回登場した際にもその単語の節点は一つしか用意しないので，すべての出現での表現の平均をその単語の表現とする．この文書グラフの節点表現を GCN を用いて更新し，その表現を使って文書分類を行う．

3.2 学習

文書グラフの節点表現を GCN で更新してその表現で文書分類を行うため，学習の時点で評価用のデータが文書グラフに含まれている必要があり，含まれていない文書についての予測はできない．そのため，学習は，分類対象のデータを利用するトランスダクティブ学習の設定で行う．具体的には，まず，評価用データのラベルを隠した上で，訓練・評価用データを両方含む文書グラフを作成し，訓練データのラベルを利用して学習することで節点の表現を更新する．学習は交差エントロピーを損失関数として用いる．BertGCN では，BERT を微調整 (fine-tuning) しながら GCN を学習させるが，本研究では，簡単のため，BERT から得られた表現は固定して用いる．BERT と GCN の同時学習は今後の課題である．

3.3 予測

学習によって，評価用データの文書節点の表現が更新されているため，この表現を利用して分類を行い，評価用データに含まれる文書のラベルを予測する．

4 実験と考察

4.1 実験設定

提案手法の評価には，医学文献の要旨のデータセットである Ohsumed と映画評論のデータセットである Movie Review を使用した．Ohsumed ではそれぞれの文書に 23 種類の心血管系疾患のカテゴリのうち 1 つ以上のカテゴリを付与されている．本研究では，単一ラベルの文書の分類にのみ注目しているので，既存研究 [2] と同様に複数ラベルを持つデータを除外した．この結果，訓練，評価用データはそれぞれ 3,357 件，4,043 件となった．Movie Review は肯定的な評論 5,331 件と否定的な評論 5,331 件を含む 2 値分類のデータセットで，訓練，評価用データはそれぞれ 7,108 件，3,554 件であった．実装には，Python 3.8.8 を使い，事前学習モデルを使うために Transformers 4.5.1[5]，モデルの作成のために Pytorch 1.8.0[6]，DGL 0.7.2[7] を用いた．評価においては，訓練データを 5 分割交差検証をして 5 つのモデルを作成した．それぞれのモデルで評価用データの予測

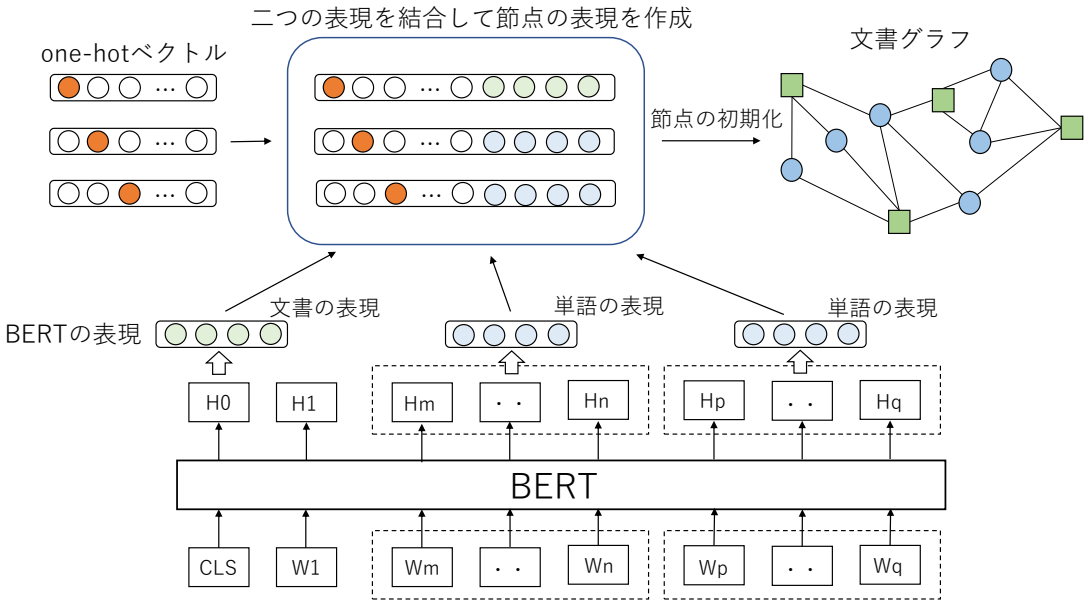

図 2　文書グラフの作成

をして，その評価の平均を最終的な予測結果として報告する．評価指標には Accuracy を用いた．最適化手法には Adam[8] を使用した．

BERT は，Movie Review には BERT-base-uncased[1] を，Ohsumed には科学文献で事前学習をしている SciBERT[9] を用いた．

ベースラインモデルとして，BERT のみを利用した文書分類モデル (BERT) と文書グラフのみを利用した文書分類モデル (GCN) を用意した．BERT のみを使用した文書分類モデルとしては，BERT でテキスト情報をエンコードして得られた出力の [CLS] トークンの表現を全結合層により予測するモデルを用意した．一般的に，BERT などの事前学習モデルは微調整して，タスクに特化させることで高い性能が達成できると知られている．しかし，本研究の提案手法ではテキスト情報を BERT に入力して，その出力を文書グラフの初期表現としているだけで，BERT を微調整せずに使っている．このことから，公平な比較を行うため，このベースラインにおいても同様に BERT は微調整せずに全結合層のみを学習することとした．文書グラフのみを利用したベースラインでは，TextGCN に倣い，節点表現を one-hot ベクトルで初期化した文書グラフを使って文書分類するモデルを用いた．

4.2　結果

実験結果を表 1 に示す．文書節点のみを BERT で初期化した文書グラフを使用したときに，Ohsumed については正解率は低下したものの，Movie Review については文書グラフを使用したベースラインよりも高い正解率を達成できた．また，BERT の文書情報のみを用いたベースラインと比較しても精度の改善が見られた．さらに，単語節点も BERT で初期化した文書グラフを使用したときには両方で正解率の改善が見られた．

単語節点の初期化については，提案手法の文書中に出現した単語の表現を利用する方法と単語の表層だけを BERT を通して表現を得る手法で比較を行った．単語の表層のみを使用する方法では，その単語のみを BERT に入力してそのサブワードの表現の平均を単語節点の初期表現とした．Ohsumed，Movie Review どちらのデータセットでも表層だけを用いた手法よりも提案手法の文章中に出現した単語の表現を用いた手法が高い正解率を示した．これは，文章中に出現した単語の文脈情報が文書グラフを用いた分類において，有用であることを示している．

5　単語節点の影響の調査

提案手法では，文書に登場するすべての単語を節点として持つ文書グラフを使用した．文書グラフを利用した文書分類においては，単語節点が文書節点の間をつなぐため，単語節点が文書間の関係を表現できることが重要である．このことから，単語節点の選択が性能にどのように寄与するかの解析を行った．ここでは，文書が単語を共有する割合が重要であると考え，単語が共通して現れる文書の数に着目

表 1　文書分類の正解率 (%)．Ohsumed には SciBERT を，Movie Review (MR) には BERT を用いた．

モデルの種類	BERT で初期化する節点	Ohsumed	MR
BERT	–	44.79	76.98
GCN	–	67.65	76.42
BERT+GCN	文書節点	65.13	77.29
BERT+GCN	文書節点・単語節点（表層のみ）	67.80	77.69
BERT+GCN	文書節点・単語節点	**68.61**	**78.78**

した．このため，それぞれの単語についてその単語が現れる文書の数を数え，その数によって，その単語を文書グラフの節点とするかどうかを閾値によって決定した．その閾値を変えて Ohsumed を用いて解析した結果を，図 3，図 4 に示す．それぞれ，節点の表現を one-hot ベクトルで初期化した文書グラフ・文書節点のみを BERT で初期化した文書グラフでの結果である．横軸が節点とするかどうかの登場回数の下限の閾値，縦軸が上限の閾値となる．図に示す数値は 5 分割交差検証における訓練データ上での正解率である．

図 3，図 4 の両方で，2 個以上 1,000 個以下の文書に登場する単語のみを節点としたときに最も高い精度となった．1 つの文書にしか登場しない単語がネガティブな影響を与えているのは，文書グラフを利用した文書分類では，単語節点が 2 つの文書をつなぐため文書間の関係を考慮した文書分類ができるが，1 つの文書としかつながらない単語節点からは文書間の関係を表現することができないためであると考えられる．また，多くの文書で登場する単語についても，多くの文書がその単語によってつながれてしまうため，カテゴリの異なる文書が近くなってしまい，正解率が下がってしまったのではないかと考えられる．

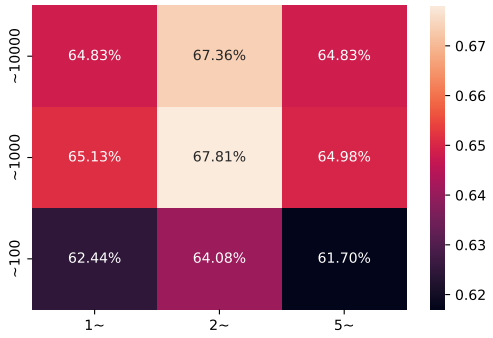

図 3　単語節点の数による正解率の比較

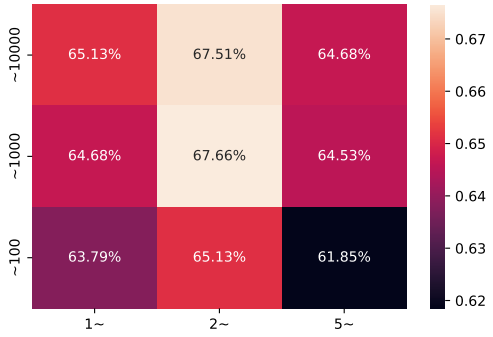

図 4　単語節点の数による精度比較

6　結論

本研究では，文書情報と文書グラフの両方を効果的に利用できる文書分類モデルを目指し，文書グラフの文書・単語の節点の両方を BERT で初期化して，GCN を学習することで，文書分類を行う手法を提案した．Ohsumed, Movie Review の 2 つのデータセットに対して評価を行った結果として，文書節点の初期値として BERT の表現を用いることで Ohsumed データについては性能が低下したものの，Movie Review データについては性能の向上が得られた．単語節点を BERT で初期化した場合については両方のデータで性能の向上が得られた．また，単語節点を BERT で初期化する際は表層よりも文脈を含めた表現を用いるほうが良いことがわかった．さらに，単語節点を対象に，文書グラフの作り方についても評価を行い，文書グラフを利用した文書分類において，1 つの文書としかつながらない単語節点や多くの文書とつながる単語節点は性能に悪影響をを与えること示した．

今後は，単語だけでなく言及や用語などの異なる要素や異なる構成を取り入れた文書グラフなど，より文書分類に適した文書グラフの作り方について検討を行う．また，BERT の微調整を GCN と同時に行う手法についても検討を行い，より効果的な文書情報と文書グラフ情報の利用方法を模索していきたい．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. **NAACL-HLT**, 2019.

[2] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 33, pp. 7370–7377, 2019.

[3] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. BertGCN: Transductive text classification by combining GNN and BERT. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1456–1462, Online, August 2021. Association for Computational Linguistics.

[4] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In **International Conference on Learning Representations (ICLR)**, 2017.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, Vol. 32, pp. 8026–8037, 2019.

[7] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. **arXiv preprint arXiv:1909.01315**, 2019.

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. **arXiv preprint arXiv:1903.10676**, 2019.

[10] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**, pp. 2623–2631, 2019.

[11] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, Vol. 15, No. 1, pp. 1929–1958, 2014.

付録

A　チューニングの詳細

Adam の学習率，GCN の隠れ層の次元数，ドロップアウト率の三個のハイパーパラメータを Optuna[10] を用いてチューニングした．ドロップアウト [11] は訓練データでの過学習を防ぐために二層の GCN の層の間に加えた．チューニングは複数のマシンで行い，NVIDIA 社の TITAN V，RTX A6000 を用いた．チューニングによって得られたハイパーパラメタは表 2 に示す．

表 2　チューニングの結果

データセット	モデルの種類	BERT で初期化する節点	学習率	隠れ層の次元数	ドロップアウト率
Ohsumed	GCN	–	0.0044	258	0.28
Ohsumed	BERT+GCN	文書節点	0.00013	103	0.70
Ohsumed	BERT+GCN	文書節点・単語節点（表層のみ）	0.0022	104	0.69
Ohsumed	BERT+GCN	文書節点・単語節点	0.00092	118	0.65
MR	GCN	–	0.0022	235	0.26
MR	BERT+GCN	文書節点	0.00036	102	0.50
MR	BERT+GCN	文書節点・単語節点（表層のみ）	0.00025	128	0.56
MR	BERT+GCN	文書節点・単語節点	0.00025	196	0.47